From Masks to Images: Mask-Conditional Diffusion for Realistic Synthetic SSS Data *

Franziska Auer[1][0009−0004−6224−776X], Zayd Yacouti[1][0009−0008−9124−6802], and Tobias Meisen[2][0000−0002−1969−559X]

[1] PTI, TKMS ATLAS ELEKTRONIK GmbH, Sebaldsbruecker Heerstrasse 235, 28309 BREMEN, Germany
{Franziska.Auer, Zayd.Yacouti}@tkmsgroup.com
[2] TMDT, University of Wuppertal, Lise-Meitner-Straße 27-31, 42119 Wuppertal, Germany
meisen@uni-wuppertal.de

Abstract. Detecting and classifying underwater objects in side scan sonar (SSS) imagery is essential for maritime safety and unexploded ordnance (UXO) remediation, yet the scarcity of labeled UXO type samples hampers modern deep learning approaches. We therefore design a modular synthesis pipeline that mirrors the physics of SSS. An unsupervised K-means segmentation generates highlight-shadow masks, the most informative cue in SSS. A mask conditional denoising diffusion probabilistic model (DDPM) generates realistic foreground objects while being forced to follow the given geometry. Simultaneously, lightweight DCGAN supplies diverse seabed textures, which are computationally cheaper than a full acoustic simulation. A physics based compositor integrates the generated foregrounds with either synthetic or real backgrounds, computes acoustic shadows analytically, and outputs YOLO style annotations. Expanding the proprietary IRAV dataset (6 classes, 20 images per class) yields 600 synthetic samples. The mask conditional DDPM achieves a mean FID of 2.31, IS of 1.07, and PSNR of 14.8 dB, confirming high visual fidelity. Augmenting a severely limited training set (60 real images, which yields a 29% test accuracy) with 600 synthetic samples lifts median test accuracy across six CNNs: Compact/medium models (MobileNetV2, ResNet-18, ResNet-50, DenseNet-121) reach an average test accuracy of 79% with real background augmentation, slightly lower than the 80% obtained with GAN backgrounds. High-capacity VGG-16/-19 benefit more from the smoother GAN textures, indicating that less complex background representation helps deep models converge better. By explicitly encoding geometry, leveraging diffusion for appearance, and using a controllable background generator, the pipeline provides a data efficient solution for SSS object classification under extreme label scarcity.

Keywords: Synthetic Images · SONAR · Mask conditional DDPM

* Supported by the BMWE (Bundesministerium für Wirtschaft und Energie) within the IRAV 2 project (Industrielle Räumung von Altlasten in Verklappungsgebieten)

1 Introduction

The detection and classification of underwater objects is a prerequisite for maritime safety, environmental monitoring, and marine resource management. Side scan sonar (SSS) is a commonly used sensor for seabed mapping because it can operate in turbid water and over large areas, yet SSS imagery is intrinsically ambiguous. Variable seafloor backscatter, sensor- and geometry-dependent acoustic shadows, and the great diversity of object shapes and material properties together make automated interpretation difficult. In legacy object rich environments such as unexploded ordnance (UXO) fields, manual identification is costly and time consuming, motivating data driven classifiers.

Public SSS datasets are growing, but few contain a balanced set of UXO-type targets. Most collections focus on wrecks, pipelines, or shipwrecks ([6], [2], [21]), leaving only a small number of images of the small, metallic objects (barrels, boxes, etc.) that dominate UXO surveys. Our proprietary IRAV dataset was recorded within a dump site providing real SSS images of UXO on the seabed while the amount of objects (the smallest class features only 20 instances) is insufficient for modern deep networks.

Consequently, our work tackles the entire target background ecosystem. We focus our investigation on two concrete research questions (RQs). The first (RQ 1) concerns generative quality: we ask how well the mask conditional Denoising Diffusion Probabilistic Model (DDPM) can reproduce both the visual appearance and the underlying statistical characteristics of SSS images. The second (RQ 2) addresses downstream usefulness: we examine whether augmenting the limited real sonar data with synthetic images leads to a measurable improvement in object classification performance, and we further explore whether the choice of background (synthetic GAN generated vs. real seabed) influences the benefit of synthetic data. To answer these questions we built an end-to-end pipeline that expands a small real sonar dataset into a big source of labeled training images.

The pipeline starts by acquiring the IRAV data, dynamically compressing its intensity range, normalizing the intensities, and pairing each image with expert generated rasterized highlight-shadow masks. To create acoustic shadow, background, and specular highlight, which are the fundamental cues in SSS imagery, we generate conditioning masks automatically by applying unsupervised K-means segmentation (k=3) to object crops. Two complementary generative models are then employed. A deep convolutional generative adversarial network (DCGAN) creates unconditional seabed textures, while a mask conditional DDPM synthesizes foreground objects under explicit guidance from the binary highlight-shadow masks. Physics-informed compositing uses simple sonar geometry to compute shadow length and acoustic shadow darkening, and blends the generated foregrounds with either a synthetic or a real background, producing fully labelled PNG images (examples given in Section 5) together with YOLO-style annotations.

Our contributions are integrated with the research questions: (1) we introduce the mask conditional DDPM for SSS, (2) we systematically compare synthetic (GAN) versus real backgrounds and quantify their impact on downstream clas-

sification, and (3) we provide an end-to-end empirical evaluation that shows how synthetic data, generated with our pipeline, close the performance gap caused by severe label scarcity.

The remainder of the paper is organized as follows. Section 2 reviews related work on acoustic image synthesis and existing SSS datasets. Section 3 details the IRAV dataset, preprocessing steps, and the architecture of each generative component. Section 4 describes the experimental setup, evaluation metrics, and the three training regimes under investigation. Section 5 presents the evaluation metrics, image quality results, and classification experiments across the three training regimes.

2 Related Work

The generation of synthetic SSS imagery has become an important strategy to mitigate the scarcity and acquisition cost of labeled real world data. Existing work can broadly be grouped into physics based simulation methods [22], [8] and data driven generative models, many of which are explicitly tailored to SSS data [6], [23].

On the simulation side, Jane et al. introduced an Unreal Engine based SSS image simulator that explicitly models acoustic phenomena such as backscatter variability and highlight-shadow patterns under different seabed conditions for automatic target recognition (ATR) [22]. Nambiar et al. presented the S3Simulator dataset, where CAD models of ships and aircraft are placed in virtual underwater environments and rendered into large collections of synthetic SSS images through a sonar specific simulation and image processing chain [8]. Woock and Beyerer proposed a physically based SSS simulator that models wave propagation, scattering, and beam patterns to generate realistic sonar images from 3D scene descriptions [28]. While these approaches capture many aspects of sidescan imaging, they require full 3D models and complex scene setups.

Data driven generative models have been explored to increase realism and variability. Steiniger et al. demonstrate transfer learning in generative adversarial networks (GANs) to generate synthetic SSS snippets, pre-training on simulated images before fine-tuning on real data [23]. More recently, conditional GANs have also been applied to specific applications, such as generating sonar images of underwater dam defects [9].

Furthermore, recent work has leveraged semisynthetic training data for underwater object detection and classification tasks in SSS. Natarajan et al. propose Synth-SONAR, a dual-stage diffusion framework that synthesizes high-resolution sonar images conditioned on text prompts using style injection and GPT-based prompting [12]. Deep transfer learning approaches using semisynthetic data have shown improved (4% increase) performance for underwater object classification [6], while multi-modal multi-stage frameworks utilize synthetic images for enhanced target recognition [27].

In contrast to methods that depend on full-scale 3-D scene simulation or that learn the sonar imaging process solely from data, our pipeline generates

SSS images directly from a compact, physically interpretable parameter set: a highlight mask, the object height, and the AUV flight altitude. The mask supplies the spatial layout of the specular highlight, while a lightweight geometric generator builds the corresponding object footprint and computes the shadow analytically with a line-of-sight model that respects the current sensor height and the object's position in the image. By encoding geometry explicitly rather than learning it implicitly, the method delivers high fidelity, consistently shaded sonar images that are ideal for large scale data augmentation and downstream classification experiments.

3 SSS generation pipeline

We present an end-to-end pipeline that turns a limited real SSS dataset into fully labeled training scenes.

3.1 Data Acquisition and Preprocessing

To obtain a balanced set of UXO targets, the proprietary IRAV dataset was recorded at a dumpsite in the Baltic Sea [15] using an Edgetech 2205 SSS [3]. All objects were classified according to three geometric descriptors (height, width and length) into six categories. Objects span 0.1m-1.4 m in size. As sonar intensity drifts across the swath, a median-filter gain correction is applied to stabilise the back scatter profile. Normalizing each frame to $[0, 1]$ and clipping to a fixed dB window ($-5\,\mathrm{dB}$ to $+20\,\mathrm{dB}$) ensures that the downstream generative models see a consistent dynamic range, which is essential for stable GAN/DDPM training. Figure 1 shows an example crop of each object class within the dataset.

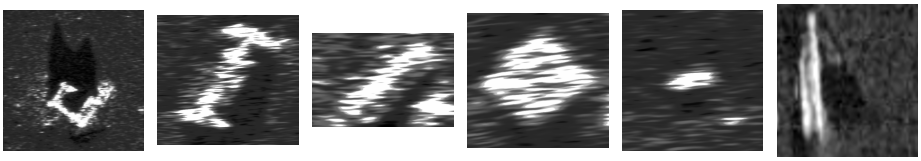

Fig. 1. Original images (left to right: Barrel, Big Elongated Box, Elongated Box, Square Box, Small targets, Distinct Objects)

We follow Steiniger et al. [24] and use 20 images per class, the maximum allowed by the smallest class.

3.2 Generation Time Mask Preparation via K-Means Segmentation

Our diffusion model requires spatial conditioning highlight-shadow masks but no manual annotations are available. We therefore generate masks automatically from the IRAV dataset using an unsupervised K-means (k=3) [1] clustering. The

15 independent restarts and a $\Delta\mu < 0.5$ stopping criterion increase robustness to speckle noise. The three intensity-sorted clusters (shadow, background, highlight) are then refined via: (1) intensity thresholding at the 80th percentile; (2) morphological opening and closing with a 3×3 cross-shaped structuring element; (3) connected component filtering (shadow > 800 px, highlight > 500 px); and (4) contrast validation, suppressing the shadow mask if the shadow/highlight intensity ratio exceeds 0.85. Validated pairs are rendered as 400×200 px grayscale PNGs.

3.3 Generative Models

To generate a large variety of 512×512 seafloor backgrounds, a DCGAN [16] was used. The generator maps a 256-d latent $\mathbf{z}\sim\mathcal{N}(0, I)$ through five transposed convolution blocks. The final activation is a `tanh` that yields values in $[-1, 1]$, later rescaled to $[0, 255]$. The discriminator mirrors this with five strided convolution blocks but accepts a 1-channel image. Training follows the standard minimax objective [4]. Both networks use Adam [10] (β_1=0.5, lr 2×10^{-4}), batch size 16, and real-sample label smoothing of 0.9. Training runs for 2000 epochs.

To implement the mask conditioning, a 2-channel binary mask $\mathbf{c}$ (separate highlight and shadow channels) is concatenated with the 3-channel pseudo-colour representation of the noisy image $\mathbf{x}_t$, obtained by applying a colourmap to the normalised single-channel sonar intensity. This early fusion strategy [13] injects the spatial highlight-shadow layout as a geometric prior at every encoder level of the U-Net backbone [19](Table 1).

Table 1. U-Net architecture for the mask conditional diffusion model.

Parameter	**Value**
Input resolution	128 × 192 px
Input channels	5 (3 RGB + 2 mask)
Output channels	3 (RGB noise prediction)
Channel schedule	128, 128, 256, 256, 512, 512
Encoder	DownBlock2D ×4, AttnDownBlock2D, DownBlock2D
Decoder	UpBlock2D, AttnUpBlock2D, UpBlock2D ×4
Total parameters	≈113M

To prevent background pixels from interfering with object appearance learning, we adopt a masked noise-prediction loss, restricting optimisation to foreground regions, inspired by spatially conditioned diffusion approaches [18]. Let $\hat{\mathbf{E}} = \hat{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c})$ denote the full noise prediction map output by the U-Net. The loss is then:

$$\mathcal{L} = \frac{\sum_{i,j} \left(\epsilon_{ij} - \hat{\mathbf{E}}_{ij}\right)^2 \cdot m_{ij}}{\sum_{i,j} m_{ij} + \delta} \tag{1}$$

where ϵ_{ij} is the ground-truth noise at pixel (i, j), $\hat{\mathbf{E}}_{ij}$ is the corresponding noise predicted by the U-Net, $m_{ij} \in \{0, 1\}$ is the binary foreground mask restricting the loss to highlight and shadow pixels, and $\delta > 0$ is a small constant for numerical stability. Training runs for 500 epochs (AdamW [11], lr 10^{-5}, batch 12), using 40% of crop-mask pairs.

A stochastic four stage augmentation pipeline is applied at inference time: (1) centroid anchored rotation (p=0.75, $\theta \in [-15°, +15°]$); (2) thin plate spline elastic deformation (p=0.70, $\sigma \in [6, 12]$ px, amplitude $\alpha \in [1.5, 3.5]$ px); (3) morphological micro perturbation (p=0.60, erosion or dilation with 3×3 kernel); (4) affine jitter (p=0.60, translation ±3 px, shear ±0.03 rad).

3.4 Physically Grounded Scene Compositor

The compositing stage must respect sonar physics so that synthetic scenes exhibit realistic shadow lengths and intensity attenuation. By computing the local grazing angle from a platform at altitude H (m) and slant range r (m) we can analytically estimate shadow length and intensity, which is essential for training detectors that rely on shadow cues. A reference grazing angle $\alpha_{\text{ref}} \sim \mathcal{U}(20°, 40°)$ is drawn per scene and used to estimate $H = r_{\text{ref}} \cdot \tan \alpha_{\text{ref}}$. Shadow length is approximated based on the object height h (m) and the pixel scale factor k (px/m) as:

$$L_s^{\text{px}} = \frac{h \cdot r}{H} \cdot k, \qquad k = 20 \text{ px/m} \tag{2}$$

Shadows are generated by ray casting from object contour pixels along $\hat{\mathbf{d}} = (1, 0)^\top$, assuming a fixed sensor-to-object geometry consistent with the IRAV survey configuration, where range increases from left to right in image coordinates. We empirically model shadow darkening as:

$$B'_{ij} = B_{ij} \cdot (1 - 0.8\, \tilde{s}_{ij}) \tag{3}$$

where B_{ij} is the original background pixel intensity at location (i, j), B'_{ij} is the resulting shadow-darkened background value, and $\tilde{s}_{ij}$ is the Gaussian-blurred (σ=7 px) binary shadow mask modelling the penumbra at the shadow boundary caused by diffraction and finite aperture effects..

Objects are sorted near-to-far and composited using alpha blending [14]:

$$O_{ij} = F_{ij} \cdot \alpha_{ij} + B_{ij} \cdot (1 - \alpha_{ij}) \tag{4}$$

where B_{ij} denotes the background pixel intensity, F_{ij} the foreground object pixel generated by the diffusion model, and O_{ij} the final composited output pixel at location (i, j). The blending coefficient $\alpha_{ij} \in [0, 1]$ represents the soft foreground mask, obtained from the union of highlight and shadow regions with edge feathering applied.

A minimum 50 px separation constraint is enforced between objects. YOLO-format labels are produced automatically. The same six object classes as in the original data (Figure 1) were used.

3.5 Alternative real background pipeline

While the GAN generated backgrounds provide a big variety, they can introduce a domain gap (synthetic artefacts). To evaluate the impact of that gap we also implement a pipeline that composites objects onto truly empty real scans, guaranteeing perfect background realism at the price of needing a larger collection of empty frames. A simple amplitude heuristic (class specific reflectivity factors) is used to match object brightness to the local background.

Table 2 compares the two compositing strategies in terms of realism vs. data acquisition overhead, reflecting the trade off we deliberately investigated.

Table 2. Comparison of the two compositing pipelines.

Property	**Pipeline 1 Real background**	**Pipeline 2 GAN background**
Background source	Real, without objects	DCGAN
Domain gap	Minimal	Possible (GAN artefacts)
Limitation of Background diversity	available scans	available training data
Amplitude domain	Native float64	RGB compositing

4 Experimental setup

The goal of the experimental work is two-fold. First, we assess how faithfully the mask conditional DDPM reproduces the visual and statistical characteristics of the IRAV SSS images (RQ 1) using three complementary metrics. Second, we investigate experimentally whether and how synthetic images can be used to improve object classification when the amount of real training data is severely limited, and how the background (GAN-generated vs. real seabed) influences this effect (RQ 2).

4.1 Evaluation Metrics

Generative model quality (RQ 1) is assessed using three widely recognized metrics ([17], [12], [7]) computed per object class: Fréchet Inception Distance (FID), Inception Score (IS), Peak Signal-to-Noise Ratio (PSNR).

FID [5] measures the distributional similarity between feature representations of real and generated images extracted from Inception-v3 [25]. Both sets are modeled as multivariate Gaussians with means $(\boldsymbol{\mu}_r, \boldsymbol{\mu}_g)$ and covariances $(\boldsymbol{\Sigma}_r, \boldsymbol{\Sigma}_g)$, and the Wasserstein-2 distance between them is computed as:

$$\mathrm{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\|_2^2 + \mathrm{Tr}\Big(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\,(\boldsymbol{\Sigma}_r \boldsymbol{\Sigma}_g)^{1/2}\Big) \tag{5}$$

where $\mathrm{Tr}(\cdot)$ denotes the matrix trace and lower FID values indicate that the generated distribution is closer to the real one. Given the very small per-class

sample size of 20 real images, features are extracted from an early convolutional layer of Inception-v3 (64-dimensional).

IS [20] quantifies both the quality and diversity of generated images via the KL divergence between the conditional class distribution $p(y|x)$ and the marginal class distribution $p(y)$:

$$\mathrm{IS}(G) = \exp\big(\mathbb{E}_{x\sim p_g}\left[D_{\mathrm{KL}}(p(y|x)\,\|\,p(y))\right]\big) \tag{6}$$

where $p(y|x)$ is the Inception-v3 class probability for generated image x, $p(y) = \int p(y|x)\,p_g(x)\,\mathrm{d}x$ is the marginal distribution over all generated images, and D_{KL} is the Kullback-Leibler divergence. Higher IS values indicate images that are both sharp (low entropy of $p(y|x)$) and diverse (high entropy of $p(y)$).

PSNR quantifies pixel-level similarity between a generated image $\hat{I}$ and a reference image I of N pixels with maximum intensity $I_{\max}$:

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{I_{\max}^2}{\mathrm{MSE}}\right), \quad \mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(I_i - \hat{I}_i\right)^2 \tag{7}$$

Since no ground-truth paired images exist, for each generated image $\hat{I}$ we compute the PSNR against every real image I_k in the same class and retain the maximum value:

$$\mathrm{PSNR}(\hat{I}) = \max_k \; 10\log_{10}\left(\frac{I_{\max}^2}{\mathrm{MSE}(\hat{I},\, I_k)}\right) \tag{8}$$

The per-class score is then the mean of these best-match values over all generated images. This protocol measures how closely each synthetic sample resembles its nearest real counterpart rather than enforcing pixel-level reconstruction from a fixed paired reference. Higher PSNR (in dB) indicates lower distortion relative to the best-matching real image.

Downstream classification performance is evaluated using two metrics computed on the IRAV dataset. Validation accuracy monitors generalization during training and is used exclusively for early stopping and checkpoint selection. Test accuracy is the primary performance indicator, measured on a fixed held-out set of 60 real patches that is never seen during training or validation, ensuring that any performance gain stems solely from the composition of the training data.

4.2 Implementation Details

To assess whether synthetic images can compensate for a severely limited real training set and how the background (GAN generated vs. real seabed) influences this effect (RQ 2), we evaluated six standard architectures provided by the torchvision library: MobileNetV2, ResNet-18, ResNet-50, DenseNet-121, VGG-16 and VGG-19. Each initialized with ImageNet weights and modified to accept a single channel input (by averaging the pretrained RGB kernels). The final fully connected layer was replaced by a six unit softmax corresponding to the

six IRAV object classes, providing a spectrum from lightweight (MobileNetV2) to high capacity (VGG-19) models.

Three mutually exclusive training configurations were evaluated. In the real-only regime the classifier was trained on 42 randomly drawn real IRAV patches (70% of the 60 available training images), with the remaining 18 patches (30%) reserved for validation. In the synthetic-GAN-background regime the same 42 real patches were augmented with 420 synthetic patches (70 per class) composited onto GAN generated seabed textures, with the 18 real validation patches supplemented by 180 synthetic patches (30 per class, unseen during training). In the synthetic-real-background regime the 42 real patches were again supplemented with 420 synthetic patches placed on real, unannotated SSS backgrounds, with validation following the same mixed scheme. The validation set was used exclusively for early stopping and checkpoint selection. The held-out test set comprised real patches only (10 per class), ensuring that any performance gain stems solely from the training data.

All networks used the same training pipeline: categorical cross-entropy loss with the Adam optimizer and a cosine-annealing schedule that reduced the learning rate to zero. Mini-batches were sized four for the real-only case and 16 for the mixed-data cases, with synthetic and real images shuffled together in the latter. Training ran up to 150 epochs, employing early stopping after 30 epochs without validation-accuracy improvement. The checkpoint with the highest validation accuracy was used for testing. A fixed random seed guaranteed reproducible data splits, weight initialization, and augmentations. Experiments were executed on an NVIDIA L40S GPU, 48 GB VRAM, using Python 3.11, PyTorch (2.5.1/2.7.0) and torchvision (0.20.1/0.22.0).

Three hypotheses were formulated before the experiments were run. The first hypothesis (H1) posits that augmenting the limited real training set with synthetic images will increase classification accuracy for all models, because the effective training size grows from 60 to 660 images regardless of the background source used. The second hypothesis (H2) expects the synthetic-real-background augmentation to yield a larger accuracy gain than the synthetic-GAN-background augmentation, because authentic real SSS backgrounds introduce the natural noise, texture variability, and acoustic characteristics of the imaging domain, thereby reducing the domain gap between training and test distributions. GAN generated backgrounds, while visually plausible, may still contain subtle artefacts or lack the fine-grained statistical properties of real seabed imagery. The third hypothesis (H3) predicts that the magnitude of the improvement will depend on model capacity: lightweight networks such as MobileNetV2 already generalise reasonably well with few examples, whereas high-capacity models like VGG-16 and VGG-19 suffer from severe overfitting on the 60-image real-only baseline and should therefore benefit most from synthetic augmentation. However, given their tendency to overfit, these large models may also be more sensitive to the quality and realism of the background, making the choice of background strategy particularly consequential for them.

5 Results

Our end-to-end pipeline was used to create synthetic images to enhance the original IRAV dataset. Figure 2 displays the synthetic image set created with a mask conditional DDPM algorithm. The pipeline produces six object categories in accordance to the original IRAV dataset (Figure 1): Barrel, Big Elongated Box, Elongated Box, Square Box, Small Targets, and Distinct Objects. Visual inspection confirms that the synthetic samples retain the critical structural cues of the target objects and exhibit photorealistic shading comparable to real data.

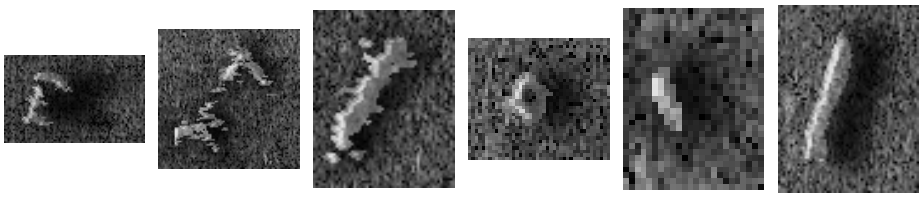

Fig. 2. Synthetic images (left to right: Barrel, Big Elongated Box, Elongated Box, Square Box, Small targets, Distinct Objects)

The experimental evidence is organized around the two research questions introduced in section 1.

5.1 Generative quality (RQ 1)

Table 3 summarises the quantitative evaluation of the mask conditional DDPM for each object class. The FID values are generally low, with a mean of 2.31 across all six classes, indicating that the synthetic samples occupy a distribution that closely matches the real sonar data. The distinct objects class achieves the smallest FID (1.13) and the highest PSNR (16.14 dB), which together suggest that the model reproduces both the visual texture and the underlying acoustic intensity statistics of this geometry with particular fidelity. In contrast, the elongated box class exhibits the largest FID (4.28), reflecting a higher divergence from the real data distribution. This can be attributed to the pronounced geometric variability.

The IS remains close to unity for all classes (average = 1.07). While low in absolute terms, this is expected given the limited visual diversity of single class original data and therefore does not necessarily reflect poor generative performance. In this context, FID provides a more reliable measure of quality.

Across the dataset the average PSNR of 14.8 dB demonstrates that the synthetic images preserve the acoustic intensity characteristics required for downstream sonar analysis tasks. Taken together, these metrics confirm that the mask conditional DDPM can reliably generate structurally plausible sonar images that retain both the visual appearance and statistical properties of the real data,

Table 3. Evaluation metrics per object class for the mask conditional DDPM.

Class	FID	IS	PSNR (dB)
Barrel	1.5428	1.0643	14.5578
Big Elongated Box	2.8803	1.0749	13.4882
Elongated Box	4.2755	1.0783	14.3986
Square Box	2.5525	1.0656	14.7111
Small Targets	1.4937	1.0610	15.5412
Distinct Objects	1.1333	1.0493	16.1408
Average	**2.3130**	**1.0656**	**14.8063**

thereby answering RQ 1 affirmatively. The only systematic limitation is observed for elongated objects, suggesting that an explicit geometry aware conditioning, could further reduce the distributional gap for such classes.

5.2 Down-stream usefulness (RQ 2)

The classification experiments were designed to test the three hypotheses introduced in Section 4 and answer RQ 2 based on the obtained results. All six models were trained under three conditions: (i) using only the limited real dataset (60 images), (ii) augmenting the real set with 600 synthetic images rendered onto GAN-generated seabed textures, and (iii) augmenting with the same number of synthetic images rendered onto real, unannotated SSS backgrounds. The quantitative outcomes, summarized in Table 4, allow a direct assessment of each hypothesis.

Table 4. Model comparison using only grayscale images (all accuracies given in mean±std).

Model	Params (M)	Split	Real data only	Real + Synthetic GAN background	Real + Synthetic Real background
MobileNetV2	2.23	val	0.9222 ±0.0351	0.9952 ±0.0045	0.9841 ±0.0081
		test	0.3200 ±0.0194	0.8267 ±0.0455	0.7700 ±0.0694
DenseNet-121	6.95	val	0.8111 ±0.0444	0.9467 ±0.0215	0.9805 ±0.0124
		test	0.2667 ±0.0471	0.7767 ±0.0455	0.7633 ±0.3230
ResNet-18	11.17	val	0.9444 ±0.0351	0.9842 ±0.0146	0.9793 ±0.0091
		test	0.3167 ±0.0149	**0.8333 ±0.0641**	0.8133 ±0.0499
ResNet-50	23.52	val	0.9333 ±0.0416	0.9903 ±0.0062	0.9829 ±0.0081
		test	**0.3267 ±0.0309**	0.8233 ±0.0442	**0.8167 ±0.0483**
VGG-16	134.29	val	0.3111 ±0.0754	0.8679 ±0.0372	0.6207 ±0.4038
		test	0.1633 ±0.0859	0.7833 ±0.0408	0.5933 ±0.2983
VGG-19	139.60	val	0.2667 ±0.0816	0.5661 ±0.2667	0.4610 ±0.3752
		test	0.1033 ±0.0067	0.4367 ±0.1830	0.3967 ±0.2647

H1 asserted that expanding the training pool from 60 to 660 images would raise the test accuracy of every architecture. The empirical results support this expectation. All six networks show a higher test accuracy when synthetic images are added, irrespective of the background type. In the real-only baseline the median test accuracy across the six networks is 29%. With GAN-background augmentation the median rises to 80%, and with real-background augmentation it reaches 77%. Thus, every architecture benefits from augmentation, confirming H1, though the magnitude of gain varies strongly with model capacity.

H2 proposed that real background augmentation would produce a larger accuracy gain than GAN background augmentation because authentic seabed imagery introduces the natural noise and variability of the imaging domain. The data contradicts this claim for the majority of networks. For all six models, test accuracy with GAN generated backgrounds is equal to or slightly higher than with real backgrounds. On average the accuracy with GAN backgrounds is 3% higher than the accuracy with real backgrounds (mean 80% vs. 77%). This suggests that synthetic, artifact free textures present a cleaner background signal, allowing the network to focus on the object features and treat the background as irrelevant. A behavior that is most effectively acquired by injecting additional stochastic texture variation (scatter) into the synthetic seabed. The smoother GAN backgrounds act as a form of regularization that mitigates overfitting, especially for the compact networks like MobileNetV2. As a result the assumption that "real backgrounds are always better" must be qualified. In severely data limited regimes, carefully controlled synthetic backgrounds can actually improve generalization, as already shown in other domains through domain randomization studies [26].

H3 anticipated that high capacity models, which overfit severely on the 60 image real only set, would benefit most from synthetic augmentation. The results partially confirm this expectation, but the benefit depends on the type of background. The two high capacity VGG models show the largest relative gains. VGG-16 jumps from 16% real only to 78% with GAN background augmentation. A big improvement, in line with H3. In comparison VGG-19 also supports H3, while it only shows modest improvement from 10% real only to 44% with GAN backgrounds. This suggests that overparameterised models are especially sensitive to the complexity and noise of real seabed imagery. The additional variability overwhelms learning when the overall dataset is tiny. Conversely, the compact and medium capacity networks (MobileNetV2, DenseNet-121, ResNet-18, ResNet-50) obtain similar absolute gains (≈50% test accuracy) from both augmentation strategies, indicating that they also profit substantially from synthetic data, albeit not more than the high capacity models. High model capacity alone does not guarantee a larger benefit from synthetic data. For overparameterised architectures the type of background matters. Even overparameterised architectures need substantial additional data to fully realise their potential and to "unlearn" prior biases, making the choice of background (clean GAN-generated or an explicit domain adaptation step) crucial.

The results directly answer RQ 2: Augmenting the scarce real sonar set with synthetic images yields a consistent boost in downstream object classification performance, and the magnitude of this boost depends on the type of background used during synthesis. The experimental evidence validates H1 broadly and H3 partially, while it refuted H2. Augmenting the scarce real sonar dataset with synthetic grayscale images reliably enhances classification accuracy for compact and medium capacity architectures (H1). GAN generated backgrounds yield slightly higher test accuracies than real-background compositing. Suggesting that artifact free backgrounds support the network to focus on learning object features during training (H2). The benefit of augmentation does correlate with model capacity. High capacity VGG models benefit more from GAN background augmentation, while other networks improve similarly under both augmentation types. This suggests that excessive background realism can hinder optimization under severe data constraints (H3). Consequently, synthetic augmentation, particularly with GAN based backgrounds, enhances downstream sonar object classification, with the strongest gains observed for higher capacity models.

6 Conclusion and Future Work

In this work we presented a fully automatic pipeline for generating high fidelity SSS imagery from a very limited real world dataset. By coupling a mask conditional DDPM with a lightweight DCGAN background generator and a physics based compositor, we were able to synthesize annotated training samples.

Our experiments confirm both research questions. The mask conditional DDPM faithfully reproduces the IRAV imagery (mean FID=2.31, IS $\approx$ 1.07, PSNR $\approx$ 14.8 dB). Only the elongated objects showed a modest gap, indicating that geometry aware conditioning might be beneficial. Augmenting the scarce real set (60 images) with 600 synthetic samples raises test accuracy for all six architectures, with a median gain of roughly 50% (the median test accuracy moves from 29% for the real only baseline to 80% for GAN background augmentation and to 77% for real background augmentation). When the synthetic objects are composited onto real seabed backgrounds, the four compact and medium capacity architectures (MobileNetV2, ResNet-18, ResNet-50, DenseNet-121) reach an average test accuracy of $\approx$79%, which is slightly lower than the $\approx$80% they achieve with GAN generated backgrounds. This shows that, for compact and medium capacity models, a clean synthetic background can be marginally more beneficial than a realistic one, possibly because it acts as an implicit regularizer. In contrast, the very high capacity VGG-16 and VGG-19 benefit more from the smoother GAN backgrounds, confirming that excessive background realism can exacerbate overfitting when data are limited.

The current compositor uses a simple linear shadow model and a fixed pixel-scale factor. Future versions should incorporate variable bottom topography, sound velocity profiles, and more realistic scattering to avoid unrealistic shadow lengths. Switching to a different sonar introduces a domain shift that could be mitigated by adding a platform parameter vector to the diffusion U-Net for

adaptive conditioning. Geometry aware conditioning should be explored to close the performance gap on elongated objects. Finally, deploying the synthetic data augmented classifiers on live UXO surveys will provide feedback on shadow modeling, domain shift handling, and overall pipeline efficacy. By addressing these points, classifiers can achieve >80% test accuracy with only a small number of annotated real sonar images, dramatically shortening the survey after processing time for UXO surveys.

Acknowledgements This research was supported by the BMWE (Bundesministerium für Wirtschaft und Energie) within the IRAV 2 project (Industrielle Räumung von Altlasten in Verklappungsgebieten 2). Furthermore, generative AI tools were utilized to assist with code development and text drafting. All content has been critically reviewed, edited, and remains the sole responsibility of the authors.

References

1. Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding. In: Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (2007)
2. Du, X., Sun, Y., Song, Y., Dong, L., Zhao, X.: Revealing the potential of deep learning for detecting submarine pipelines in side-scan sonar images: An investigation of pre-training datasets. Remote Sensing (2023)
3. EdgeTech: 2205: Auv / uuv / rov / asv / usv sonars. https://www.edgetech.com/product/2200-and-2205-auv-uuv-rov-asv-usv-sonars/ (2021), accessed October 28, 2025
4. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (2014)
5. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems (2017)
6. Huo, G., Wu, Z., Li, J.: Underwater object classification in sidescan sonar images using deep transfer learning and semisynthetic training data. IEEE Access (2020)
7. Husseini, S., Dugelay, J.L.: A comprehensive framework for evaluating deepfake generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), DFAD Workshop (2023)
8. Kamal Basha, S., Nambiar, A.: S3simulator: A benchmarking side scan sonar simulator dataset for underwater image analysis. In: International Conference on Pattern Recognition (2024)
9. Kang, F., Tang, P., Chen, J., Geng, Z., Zhou, D.: Synthesizing sonar image of underwater dam defects with conditional generative adversarial networks. Engineering Structures (2026)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

12. Natarajan, P., Basha, K., Nambiar, A.: Synth-sonar: Sonar image synthesis with enhanced diversity and realism via dual diffusion models and gpt prompting. arXiv preprint arXiv:2410.08612 (2024)
13. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning (ICML) (2021)
14. Porter, T., Duff, T.: Compositing digital images. In: ACM SIGGRAPH Computer Graphics (1984)
15. Projektträger Jülich (PtJ): Statustagung maritime technologien (2025), https://www.ptj.de/foerdermoeglichkeiten/maritime-forschungsstrategie/statustagung, accessed February 26, 2026
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
17. Roja, D., Jidugu, S.K., Rao, T.S., Vuyyuru, L.R., Vellela, S.S., Ranjani, B.S.: High-fidelity image synthesis using enhanced generative adversarial networks with attention mechanisms. In: 2025 International Conference on NexGen Networks and Cybernetics (IC2NC) (2025)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015)
20. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems (2016)
21. Sethuraman, A., Sheppard, A., Bagoren, O., Pinnow, C., Anderson, J., Havens, T., Skinner, K.: Machine learning for shipwreck segmentation from side scan sonar imagery: Dataset and benchmark. Journal: International Journal of Robotics Research (2024)
22. Shin, J., Chang, S., Bays, M.J., Weaver, J., Wettergren, T.A., Ferrari, S.: Synthetic sonar image simulation with various seabed conditions for automatic target recognition (2022)
23. Steiniger, Y., Kraus, D., Meisen, T.: Generating synthetic sidescan sonar snippets using transfer learning in generative adversarial networks. Journal of Marine Science and Engineering (2021)
24. Steiniger, Y., Stoppe, J., Meisen, T., Kraus, D.: Dealing with highly unbalanced sidescan sonar image datasets for deep learning classification tasks. In: Global Oceans 2020: Singapore - U.S. Gulf Coast (2020)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS) (2017)
27. Wang, J., Li, H., Huo, G., Li, C., Wei, Y.: Multi-modal multi-stage underwater sidescan sonar target recognition based on synthetic images. Remote Sensing (2023)
28. Woock, P., Beyerer, J.: Physically based sonar simulation and image generation for side-scan sonar. Uace 2017 (2017)